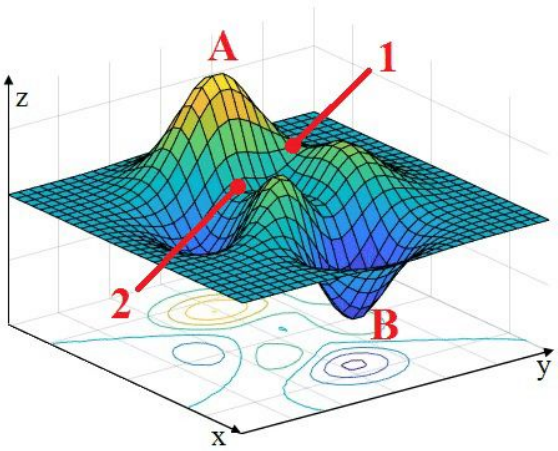
Potential Energy Prediction with Persistence Images

Owen Queen, Sai Thatigotla, Henry Eigen

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Problem Statement

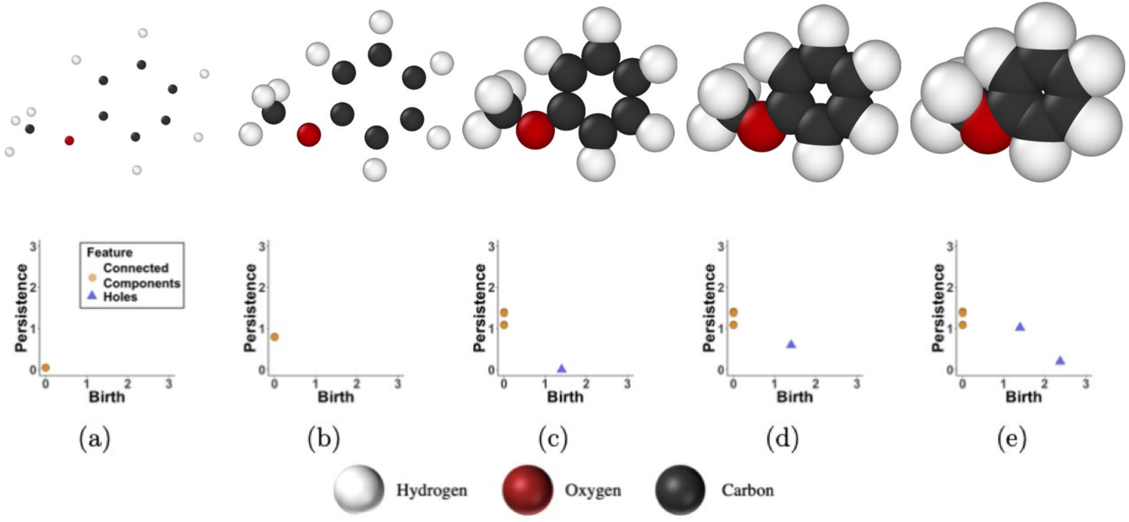
- Predict potential energy surface of a molecule
  - Ultimately, find lowest energy conformer of the molecule



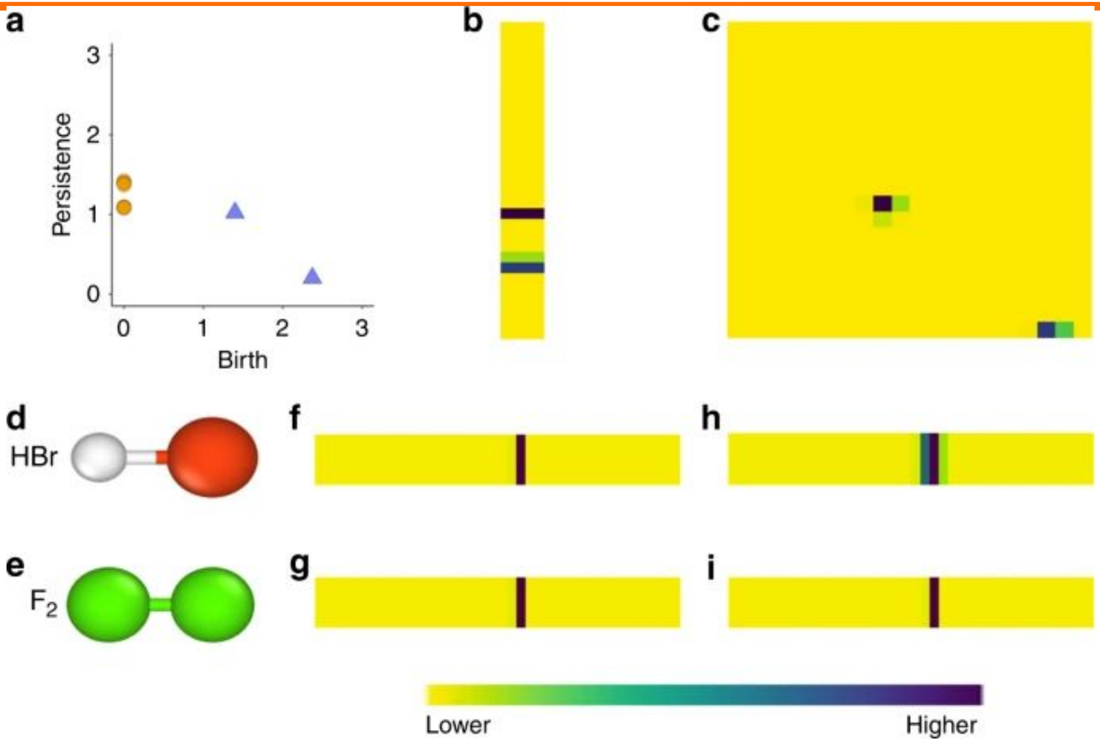
Framing the Problem with ML

- Δ geometry $\rightarrow$ Δ energy
  - How to capture geometric perturbations?
- Need to solve…
  - Regression problem
  - Optimization problem

Persistence Diagrams

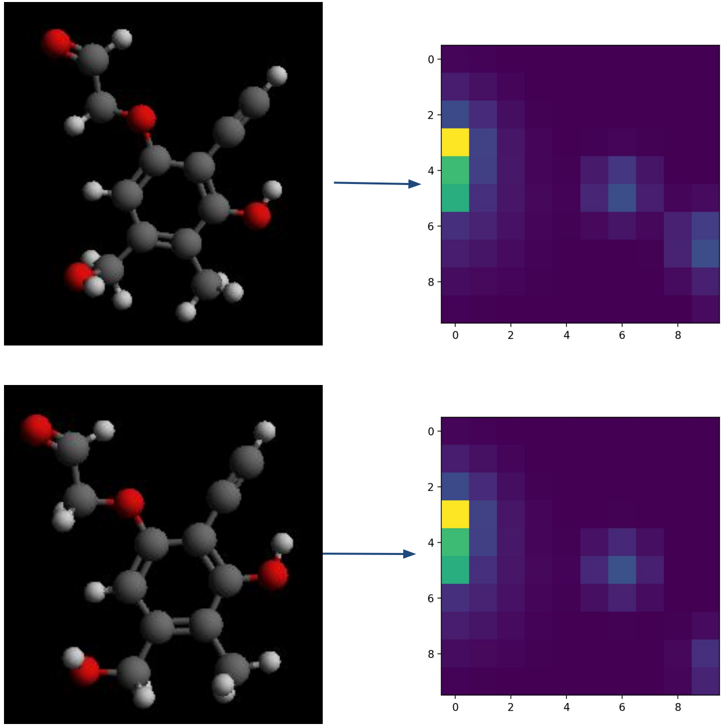


Persistence Images



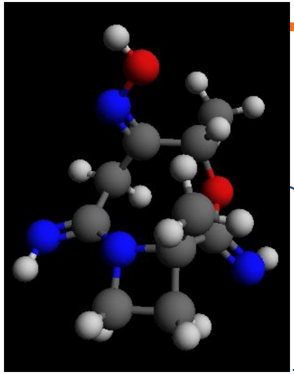
Generating Conformations

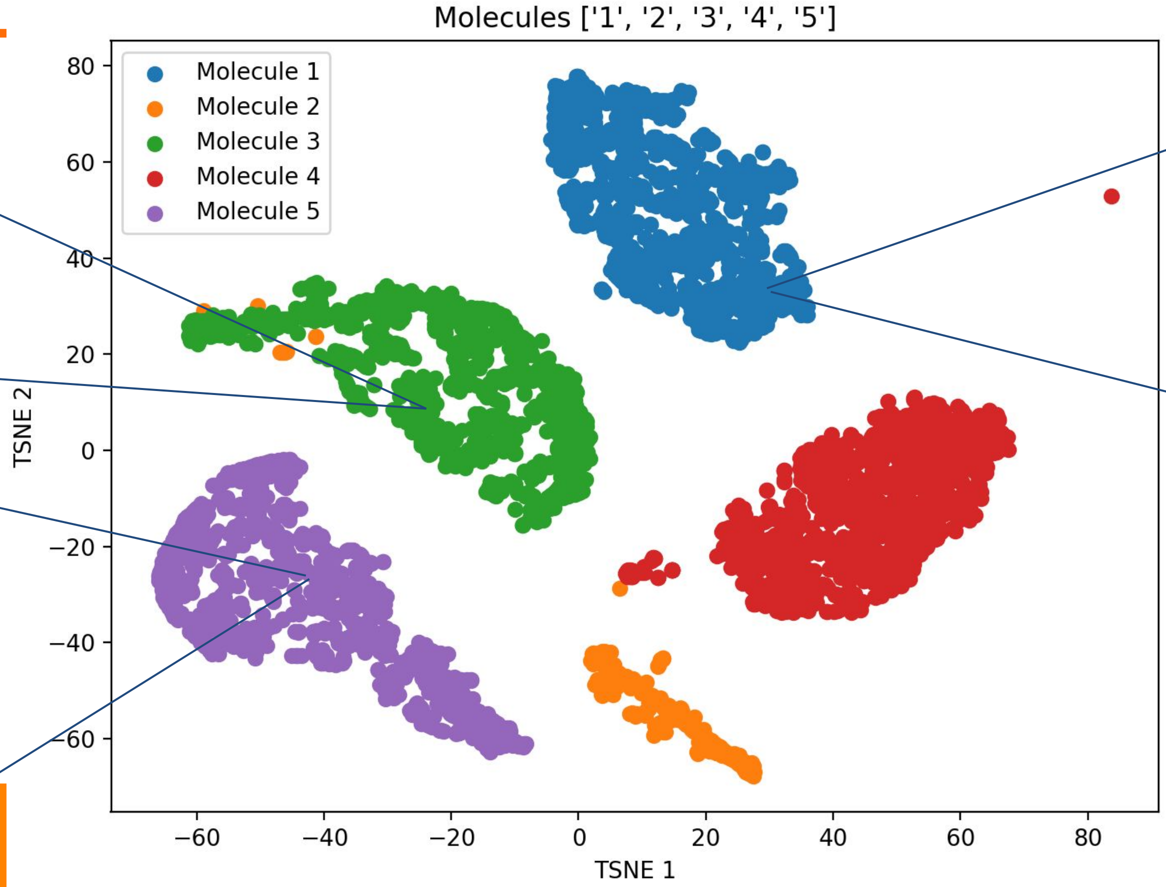
- Base structures from GDB-17 database[1]
- Conformations generated with XTB[2]
    - Uses molecular dynamics

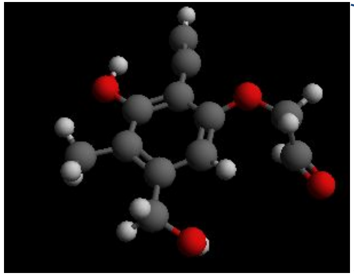


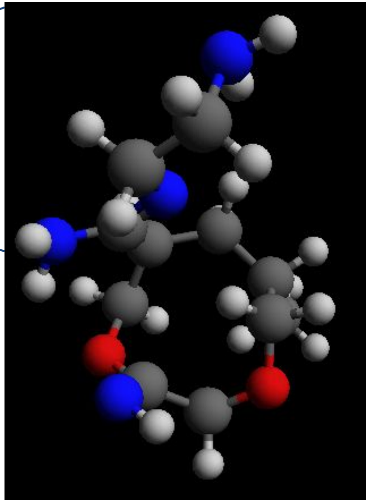
[1]L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model*, **2012**.
[2]C. Bannwarth et al., *WIREs Comp. Mol. Sci.*, **2020**.

Distribution of PI’s



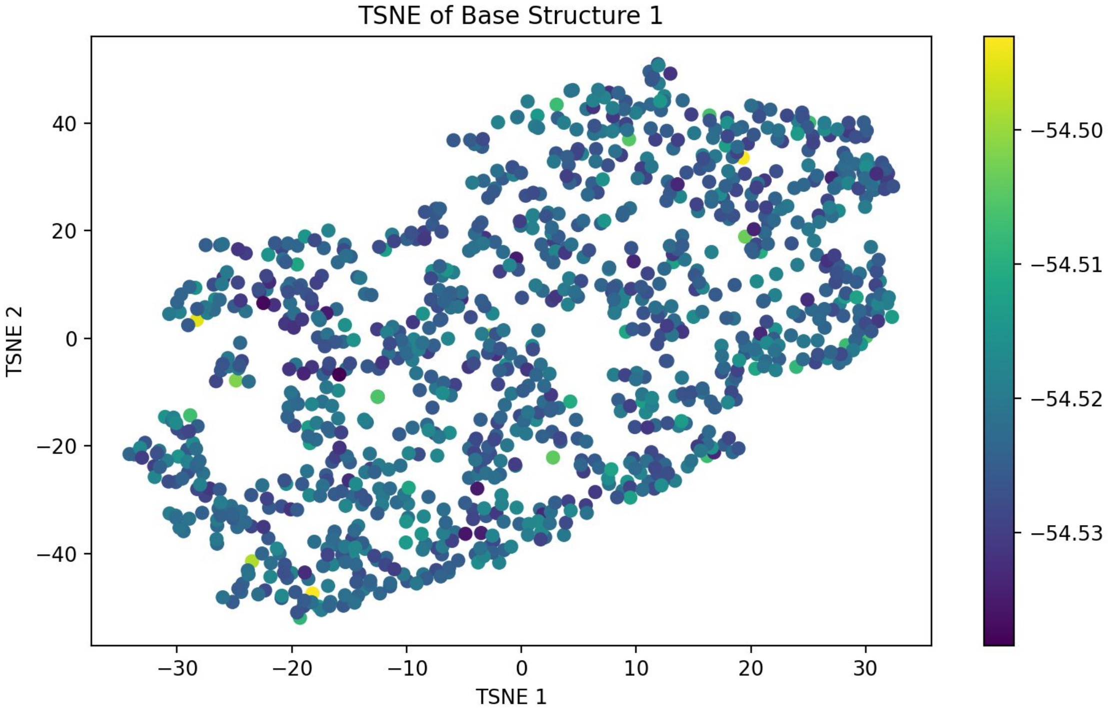




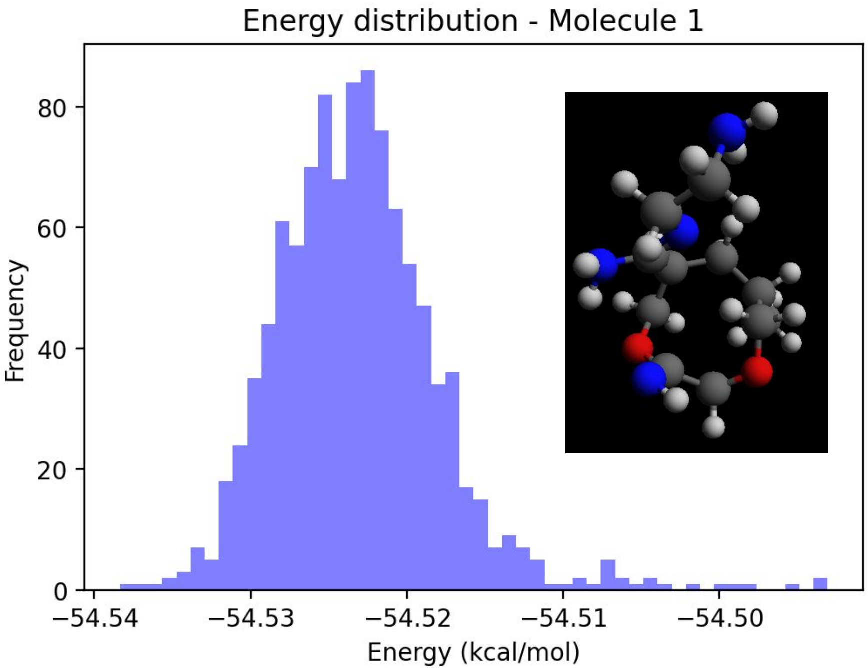


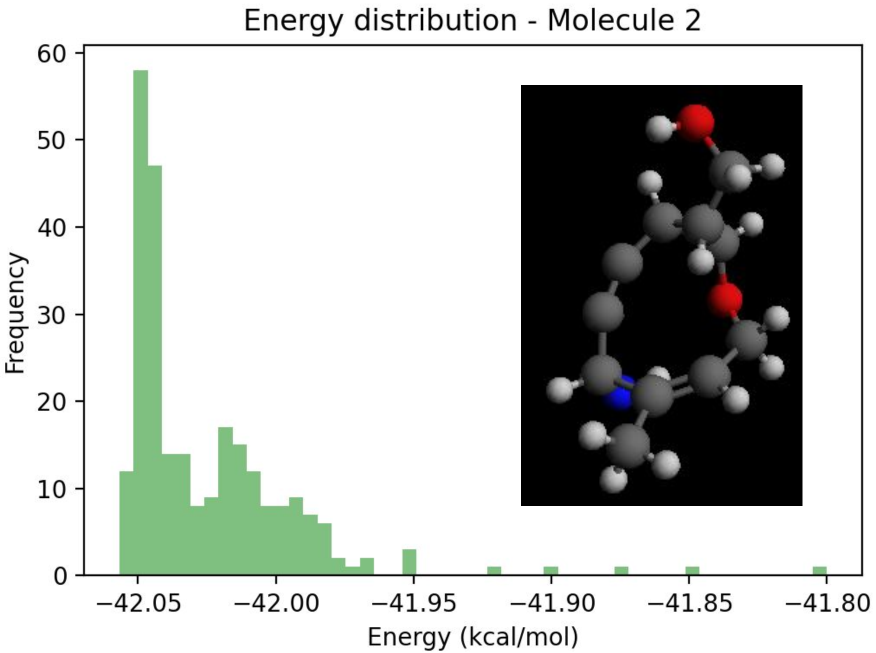
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Distribution of Energy per PI



Energy Distributions

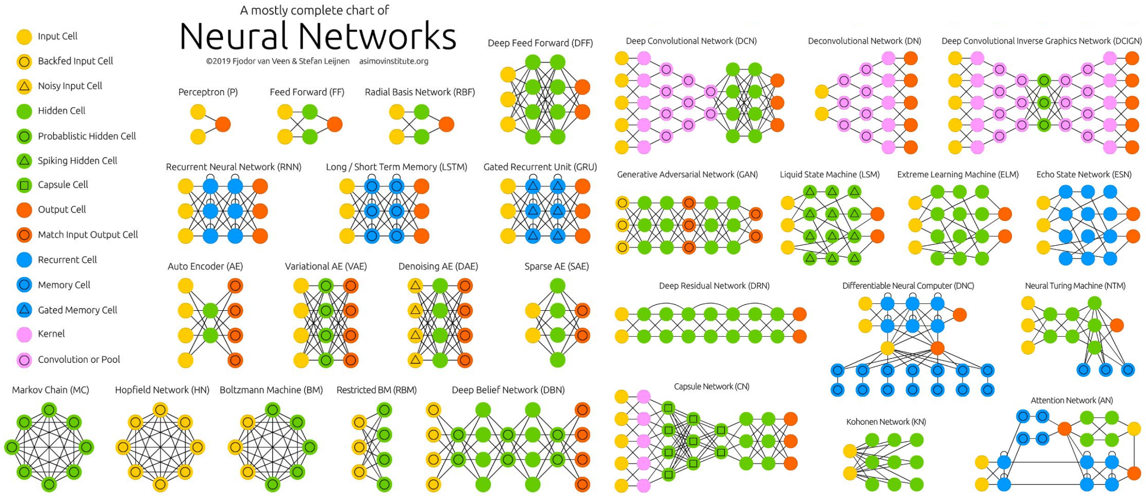




ML Models

- Non-deep learning
  - XGBoost, Kernel Ridge Regression, Support Vector Machine
- Deep learning
  - Convolutional neural network (CNN)
  - Recurrent Neural Network (RNN)

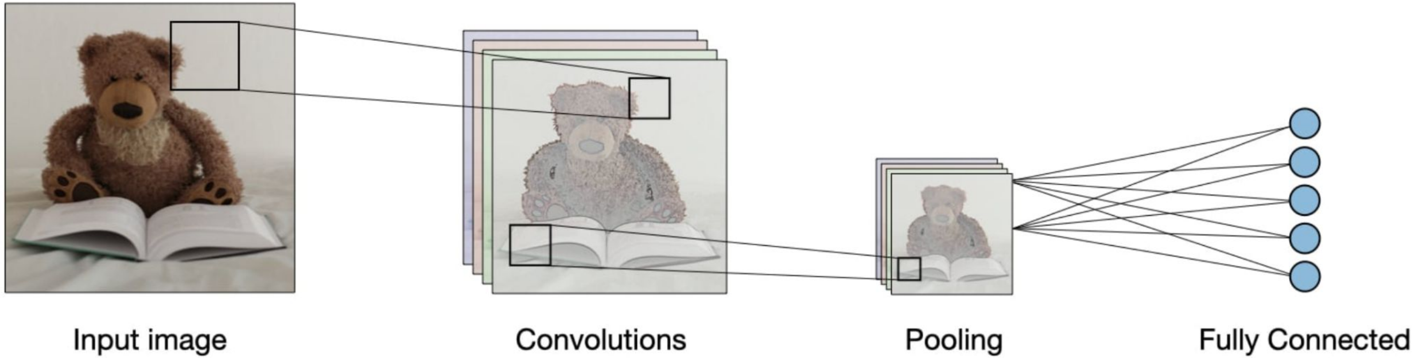
Deep Learning Architectures

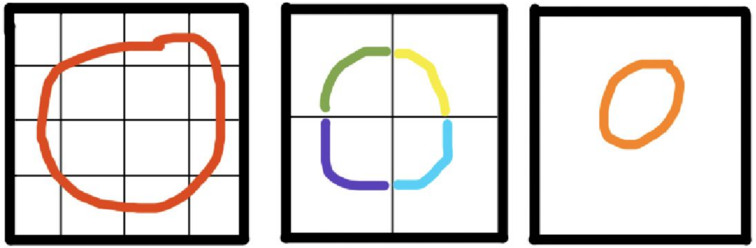


- Any architecture can learn anything (sorta)
- Goal is to find architecture that learns efficiently for given data
  - Image vs. Text
  - Sparse vs. Dense
  - Dynamic vs. Static
  - etc.

*Image from Asimov Institute

Convolutional Neural Networks

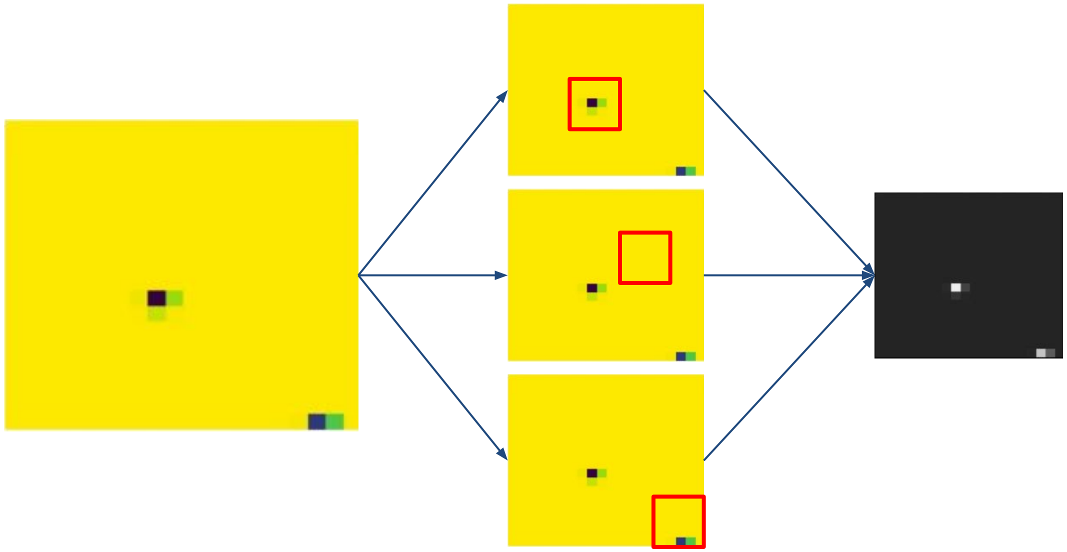




- Local Spatial Information, not global
- Combines features and downscales
- Good for learning hierarchical feature representations

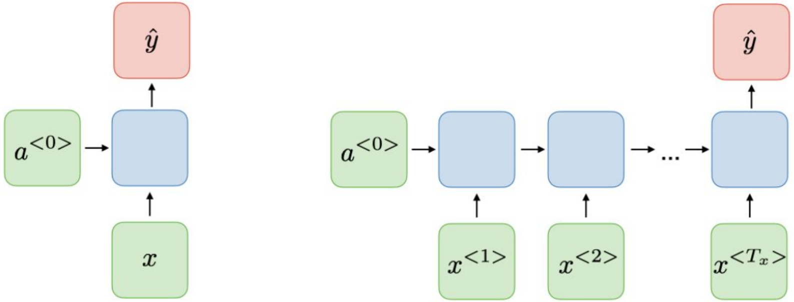
*Image from Shervine CS230 @ Stanford

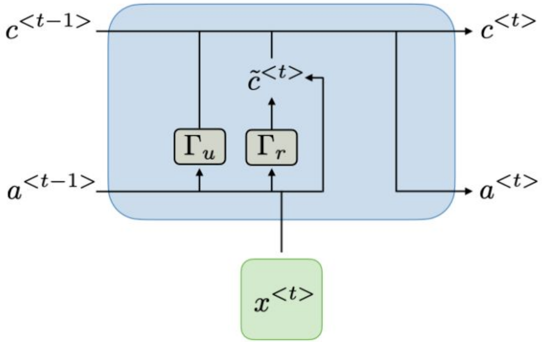
Convolutional Neural Networks Cont.



- Sparse data
- Global spatial information more important than feature shape

Recurrent Neural Networks





- How does stored information affect interpretation of new observation
- How much of past stored information should be replaced with information from new observation

*Images from Shervine CS230 @ Stanford

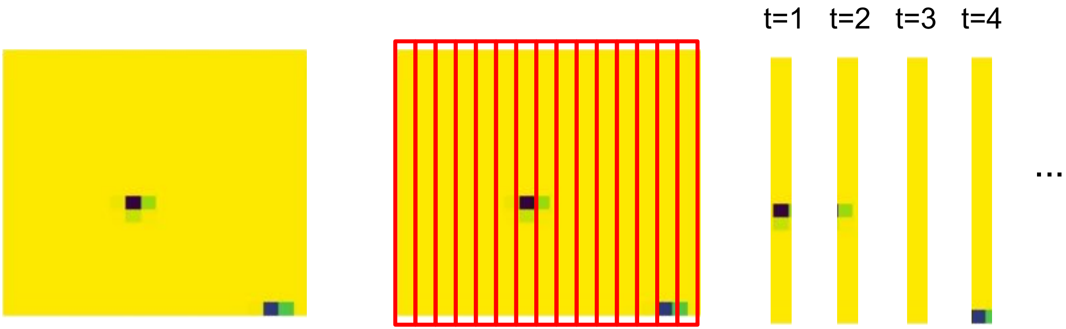
Recurrent Neural Networks Cont.

One Hot Encoding

Rome Paris word V

```
Rome   = [1, 0, 0, 0, 0, 0, ..., 0]
Paris  = [0, 1, 0, 0, 0, 0, ..., 0]
Italy  = [0, 0, 1, 0, 0, 0, ..., 0]
France = [0, 0, 0, 1, 0, 0, ..., 0]
```

PI as Time Series Data



Training Details

- Split: 70% training, 30% testing
- Results reported in Mean Absolute Error (MAE)
- Regular ML: Sci-kit Learn
- CNN: PyTorch
  - Accuracy reported on lowest test loss observed during training
  - Trained on MAE as loss function
- Only trained/tested on conformers of Base Structure 1

Results - One Structure

Non-deep learning models (Mean Absolute Error kcal/mol)

- XGBoost: 0.0033107
- Kernel Ridge Regression: 0.0069986
- Support Vector Machine: 0.0076822

Deep Learning:

- CNN: ~0.004
- RNN: still in development

Open Questions

- How well do PIs work for this problem?
  - Can they capture subtleties in geometries?
- Which ML algorithms optimally capture information in PIs?
- How to optimize for lowest-energy conformation?

Future Work

- Graph neural networks (GNNs)
- Implement recurrent neural networks (RNNs)
- Autoencoders
- Transfer learning across molecules
- Using PIs as augmentation tool